UDC: 004.8:004.032.26 DOI: https://doi.org/10.15407/jai2022.01.239

ADAPTIVE NEURO-FUZZY CLUSTERING OF DISTORTED DATA BASED ON PROTOTYPE-CENTROID STRATEGY USING EVOLUTIONARY PROCEDURES

Y. Bodyanskiy[1], I. Pliss[2], A. Shafronenko[3]
[1,2,3]Kharkiv National University of Radioelectronics, Ukraine
Nauky Ave, 14, Kharkiv, 61166
yevgeniy.bodyanskiy@nure.ua; iryna.pliss@nure.ua; alina.shafronenko@nure.ua

[1]http://orcid.org/0000-0001-5418-2143
[2]http://orcid.org/0000-0001-7918-7362
[3]http://orcid.org/0000-0002-8040-0279

Annotation. The problem of clustering is a very relevant area in Data Mining of different nature. To solve this problem, there are a large number of known methods and algorithms, most of which work in batch mode, in conditions when the entire of data set is known in advance and does not change over the time. These methods are complex in software implementation and are not without drawbacks.

The aim of the work is to develop an adaptive neuro-fuzzy clustering method of distorted data based on prototype-centroid strategy using evolutionary procedures, that solves the problem in online mode, when data are fed sequentially in real time and is characterized by numerical simplicity and high speed.

The problem of adaptive fuzzy clustering of distorted data by outliers and emissions, which are presented in the form of vector arrays, based on the strategy of the nearest prototype - centroid using evolutionary procedures, is considered. The proposed approach is based on the online probabilistic fuzzy clustering procedure with the membership function of special type and the evolutionary cat swarm algorithm.

Proposed adaptive neuro-fuzzy clustering method of distorted data based on prototype-centroid strategy using evolutionary procedures characterized by computational simplicity, high speed and accuracy of the results based on experimental studies.

The modification of optimization procedure that based on cat swarm algorithm was propose. The proposed method is simple in numerical implementation, workable in the case when the data is distorted and are fed sequentially in online mode, that is confirmed experimentally.

Keywords: evolutionary algorithm of cat swarms, prototype - centroid, adaptive fuzzy clustering.

АДАПТИВНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ВИКРИВЛЕНИХ ДАНИХ НА ОСНОВІ СТРАТЕГІЇ НАЙБЛИЖЧОГО ПРОТОТИПУ – ЦЕНТРОЇДА З ВИКОРИСТАННЯМ ЕВОЛЮЦІЙНИХ ПРОЦЕДУР

Є. В. Бодянський[1], І. П. Плісс[2], А. Ю. Шафроненко[3]
[1,2,3]Харківський національний університет радіоелектроніки, Україна
пр. Науки, 14, м. Харків, 61166
yevgeniy.bodyanskiy@nure.ua; iryna.pliss@nure.ua; alina.shafronenko@nure.ua

[1]http://orcid.org/0000-0001-5418-2143
[2]http://orcid.org/0000-0001-7918-7362
[3]http://orcid.org/0000-0002-8040-0279

Анотація. Задача кластеризації досить часто зустрічається в інтелектуальному аналізі даних різної природи. Для вирішення цієї проблеми існує велика кількість відомих методів та алгоритмів, які здебільшого працюють в пакетному режимі, в умовах, коли вся вибірка даних відома заздалегідь та не змінюється з часом. Ці методи складні в програмній реалізації та не позбавлені недоліків.

Мета роботи полягає в розробці адаптивного метода кластеризації викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур, якій вирішує задачу в онлайн-режимі, тобто коли дані надходять послідовно в реальному часі та характеризуються чисельною простотою та високою швидкодією.

Розглянуто задачу адаптивної нечіткої кластеризації викривлених збуреннями та викидами даних, які представлені у вигляді масивів векторних даних на основі стратегії найближчого прототипу - центроїда з використанням оптимізаційних процедур. В основі запропонованого підходу лежить онлайн ймовірнісна процедура нечіткої кластеризації із функцією належності спеціального вигляду та еволюційний алгоритм котячих зграй.

Особливістю запропонованого адаптивного методу кластеризації викривлених даних на основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур є обчислювальна простота, висока швидкість та точність отриманих результатів, що підтверджуються експериментальними дослідженнями.

Запропоновано модифікацію, введену на основі процедури оптимізації котячих зграй з покращеними властивостями за рахунок використання стохастичної оцінки градієнта. Запропонований метод є простим у чисельній реалізації, працездатним у випадку, коли дані пошкоджені та надходять послідовно в онлайн-режимі, що підтверджено експериментально.

Ключові слова: еволюційний алгоритм котячих зграй, прототип-центроїд, адаптивна нечітка кластеризація.

Вступ

Проблема нечіткої кластеризації викривлених даних достатньо поширена серед багатьох сфер сьогодення і є невід'ємною частиною загального напрямку обчислювального інтелекту. Для вирішення цієї задачі було запропоновано безліч методів та алгоритмів інтелектуального аналізу даних, найбільш ефективними серед яких є методи, що базуються на штучних нейронних мережах, м'яких обчисленнях тощо [1-3]. Усі ці методи працездатні лише у випадках, коли дані надходять на обробку у пакетному режимі й не змінюються з часом. Тому розробка процедур адаптивної нечіткої кластеризації викривлених даних, що вирішують задачу в онлайн-режимі, тобто коли дані надходять послідовно в реальному часі, та характеризуються чисельною простотою та високою швидкодією є актуальною.

Адаптивна нечітка кластеризація викривлених даних на основі стратегії найближчого прототипа-центроїда

Вихідною інформацією є дані, що представлені у вигляді $(N \times n)$ таблиці "об'єкт-властивість" яка містить інформацію про N об'єктів, описаних у вигляді $(1 \times n)$ векторів - ознак. Результатом кластеризації вихідних даних є розбиття початкової вибірки на m класів з відповідним рівнем нечіткої залежності $U_q(k)$ k-того вектора-спостереження до

q-го кластера, де $1 \le q \le m$. Вихідні дані заздалегідь нормуються в гіперкуб $[-1;1]^n$.

Стратегія найближчого прототипу-центроїда може бути розглянута в якості гібрида стратегії оптимального розширення та часткових відстаней і складається з послідовності кроків:

1.Завдання початкових умов для роботи методу: $\beta > 0$, m, необхідної точності $\varepsilon > 0$ прототипів (центроїдів) кластерів w_q, кількості епох $\tau = 1, 2, ..., Q$.

2.Розрахунок рівнів належності:

$$U_q^{(\tau+1)}(k) = \left(\sum_{l=1}^{m} \left(\left\| \hat{x}^{(\tau)}(k) - w_l^{(\tau)} \right\|^2 \right)^{\frac{1}{1-\beta}} \right)^{-1} \left(\left\| \hat{x}^{(\tau)}(k) - w_q^{(\tau)} \right\|^2 \right)^{\frac{1}{1-\beta}}.$$

3.Розрахунок центроїдів кластерів:

$$w_q^{(\tau+1)} = \left(\sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k) \right)^{\beta} \right)^{-1} \sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k) \right)^{\beta} \hat{x}^{(\tau)}(k).$$

4.Перевірка умов останова:

якщо $\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| < \varepsilon \ \forall \ q$

або $\tau = Q$, останов;

інакше йти до кроку 5.

5.Оцінка спотворених спостережень шляхом знаходження прототипу $w_q^{(\tau+1)}$ найближчого до $\tilde{x}(k)$ в сенсі часткової відстані

$$D_P^2\left(\tilde{x}(k), w_q\right) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} \left(\tilde{x}_i(k) - w_{qi} \right)^2 \delta_{ki},$$

тобто знаходження

$$w_q^{(\tau+1)} = \arg\min_q \left\{ D_P^2\left(\tilde{x}(k), w_1^{(\tau+1)}\right), ..., D_P^2\left(\tilde{x}(k), w_m^{(\tau+1)}\right) \right\}$$

і заміна відсутніх спостережень $\tilde{x}_i(k)$ координатами $\hat{x}_i^{(\tau+1)}(k) = w_{qi}^{(\tau+1)}$.

Далі йти до кроку 2.

Далі можна записати стратегію найближчого прототипу у рекурентній формі [...]

$$\begin{cases} U_q^{(\tau+1)}(k) = \left(\sum_{l=1}^{m} \left(\left\| \hat{x}_k^{(\tau)} - w_l(k) \right\|^2 \right)^{\frac{1}{1-\beta}} \right)^{-1} \left(\left\| \hat{x}_k^{(\tau)} - w_q(k) \right\|^2 \right)^{\frac{1}{1-\beta}}, \\ \partial e \ \hat{x}_i^{(\tau)}(k) = w_{qi}(k), \\ w_q(k) = \arg\min_q \left\{ D_P^2\left(\tilde{x}(k), w_1(k)\right), \dots, D_P^2\left(\tilde{x}(k), w_m(k)\right) \right\}, \\ w_q(k+1) = w_q(k) + \eta(k+1)\left(U_q^{(Q)}(k)\right)^{\beta} \left(\hat{x}^{(Q)}(k) - w_q(k)\right) \ \ \forall q = 1, 2, \dots, m. \end{cases}$$

Можлива стратегія найближчого прототипу-центроїда у загублених спостереженнях може бути записана у вигляді послідовності кроків:

1.Завдання початкових умов для роботи методу: $\beta > 0$, m, необхідної точності $\varepsilon > 0$ прототипів (центроїдів) кластерів w_q, кількість епох $\tau = 1, 2, \dots, Q$.

2.Розрахунок рівнів належності:

$$U_q^{(\tau+1)}(k) = \frac{1}{1 + \left(\dfrac{\left\| \hat{x}^{(\tau)}(k) - w_q^{(\tau)} \right\|^2}{\mu_q^{(\tau)}} \right)^{\frac{1}{\beta-1}}}.$$

3.Розрахунок центроїдів кластерів:

$$w_q^{(\tau+1)}(k) = \left(\frac{\sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k)\right)^{\beta} \hat{x}^{(\tau)}(k)}{\sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k)\right)^{\beta}} \right).$$

4.Перевірка умов останова: якщо $\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| < \varepsilon \ \forall \ q$ або $\tau = Q$, останов; інакше йти до шага 5.

5.Оцінка відсутніх спостережень шляхом знаходження прототипу $w_q^{(\tau+1)}$ найближчого до $\tilde{x}(k)$ в сенсі часткової відстані

$$D_P^2\left(\tilde{x}(k), w_q\right) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} \left(\tilde{x}_i(k) - w_{qi} \right)^2 \delta_{ki},$$

тобто знаходження

$$w_q^{(\tau+1)} = \arg\min_q \left\{ D_P^2\left(\tilde{x}(k), w_1^{(\tau+1)}\right), \dots, D_P^2\left(\tilde{x}(k), w_m^{(\tau+1)}\right) \right\}$$

і заміна відсутніх спостережень $\tilde{x}_i(k)$ координатами $\hat{x}_i^{(\tau+1)}(k) = w_{qi}^{(\tau+1)}$.

6.Розрахунок скалярного параметра відстані

$$\mu_q^{(\tau+1)} = \frac{\sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k)\right)^{\beta} \left\| \hat{x}^{(\tau+1)}(k) - w_q^{(\tau+1)} \right\|^2}{\sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k)\right)^{\beta}}.$$

7.Далі йти до кроку 2.

Аналогічно ймовірнісній адаптивній кластеризації на основі стратегії найближчого центроїда можна організувати процес можливісної кластеризації у вигляді [10].

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{1}{1 + \left(\dfrac{\left\| \hat{x}_k^{(\tau)} - w_q(k) \right\|^2}{\mu_q^{(\tau)}} \right)^{\frac{1}{\beta-1}}}, \\ \partial e \ \hat{x}_i^{(\tau)}(k) = w_{qi}(k), \ \ w_q(k) = \arg\min_q \left\{ D_P^2\left(\tilde{x}(k), w_1(k)\right), \dots, D_P^2\left(\tilde{x}(k), w_m(k)\right) \right\}, \\ w_q(k+1) = w_q(k) + \eta(k+1)\left(U_q^{(Q)}(k)\right)^{\beta} \left(\hat{x}^{(Q)}(k) - w_q(k)\right) \ \ \forall q = 1, 2, \dots, m, \\ \mu_q^{(\tau+1)} = \dfrac{\sum_{p=1}^{k} \left(U_q^{(\tau+1)}(p)\right)^{\beta} \left\| \hat{x}^{(\tau)}(k) - w_q(k) \right\|^2}{\sum_{p=1}^{k} \left(U_q^{(\tau+1)}(p)\right)^{\beta}}. \end{cases}$$

Оптимізаційна процедура на основі еволюційного алгоритму котячої зграї

Для знаходження локальних екстремумів у вихідних даних, що надходять на обробку методами адаптивної нечіткої кластеризації даних на основі стратегії найближчого прототипу - центроїда доцільно використовувати еволюційні алгоритми рою частинок [4-6]. Одним з найшвидших алгоритмів рою частинок є, так званий, алгоритм котячої зграї [7], який підтвердив свою ефективність у вирішенні широкого кола задач від елементарних завдань Data Mining до більш складних задач: Dynamic Data Mining, Data Stream Mining, Big Data Mining, Web Mining, Text Mining тощо.

Даний алгоритм використовує модель поведінки котів у зграї (CS), яка складається з Q особин, при цьому вважається, що кожен кіт cat_p $(p = 1, 2, \dots, Q)$ зграї може знаходитись в одному з двох положень: ре-

жим пошуку (SM), який пов'язаний із повільними рухами незначної амплітуди біля вихідної позиції або режим трасування (TM), який визначається швидкими стрибками з великою амплітудою та дозволяє вивести кота з локального екстремуму, якщо він туди потрапив. Поєднання цих станів кота дозволяє з більшою ймовірністю відшукати глобальний екстремум у порівнянні з традиційними методами багатоекстремальної оптимізації [8, 9]. У загальному випадку обидва ці режими для кожного з котів можуть бути описані процедурою оптимізації [11, 13]:

$$cat_p(\tau+1) = cat_p(\tau) - \alpha(cat_p(\tau) - cat_p(\tau-1)) - \eta\hat{\nabla}E_M(cat_p(\tau)) + \eta_\xi \Xi(\tau),$$

де $cat_p(\tau+1)$ - стан (режим) кота p на τ - ітерації, α - параметр, що визначає інерційні властивості в режимі трасування, η - крок режиму пошуку, $\hat{\nabla}E(cat_p(\tau))$ градієнтна оцінка цільової функції методу кластеризації, $\Xi(\tau)$ - випадкова компонента, яка вносить додаткові стохастичні рухи в режимі трасування, η_ξ - параметр, що визначає амплітуду цих рухів.

Цей підхід забезпечує пошук глобального екстремуму у випадку, коли кількість котів у зграї достатня.

Експериментальні дослідження

Експериментальні дослідження запропонованого методу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур було проведено на чотирьох різних вибірках даних, які були штучно пошкоджені викидами та пропусками. У таблиці 1 наведено характеристики вибірок та кількість пошкоджених даних у відсотках (%), у таблиці 2 наведені параметри для оптимізаційного методу котячих зграй (CSO).

Порівняльні експерименти запропонованого методу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур проводились з більш відомими алгоритмами кластеризації, такими як алгоритми k - середніх та k - прототипів та вимірялись за чотирма характеристиками: F-Measure, Rand Index, Jaccard Index і Ентропія. Усі ці чотири показники мають значення від 0 до 1.

Таблиця 1. Характеристики вибірок та кількість пошкоджених даних у відсотках (%)

Вибірка	Кількість кластерів	Кількість атрибутів	Кількість спостережень	Кількість викривлених даних (%)
Hepatitis	2	19	155	10
Cancer	2	9	683	50
Stat Log Heart	2	13	270	25
Post Operative Patient	6	8	214	5

Таблиця 2. Параметри для оптимізаційного методу котячих зграй (CSO)

Параметри	Значення
SRD	Випадково [0,1]
Seeking memory Pool (SMP)	5
Розмір популяції	Кількість кластерів
r_1	Випадково в діапазоні [0,1]
c_1	Const
SPC	Випадково в діапазоні [0,1]
Кількість ітерацій	Вручну

В F-Measure, Rand Index та Jaccard Index значення одиниці вказує, що кластери даних абсолютно однакові, а збільшення значень цих показників свідчить на кращу продуктивність. У таблицях 3, 4, 5 наведено результати порівняльної роботи відомих методів кластеризації даних із запропонованим методом адаптивної нечіткої кластеризації, викривлених пропусками та викидами даних основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур (AFC_PCEP). Як видно із порівняльних таблиць, запропонований метод демонструє достатньо високі показники, незалежно від вибірки та якості даних, на відміну від більш відомих методів

кластеризації даних, показник якого найближче до одиниці, що само по собі свідчить про високу якість кластеризації даних.

Таблиця 3. Порівняльні результати методів за характеристикою F-Measure

Вибірка	K-Means	K-Prototype	AFC_PCEP
Hepatitis	0.75	0.86	0.88
Cancer	0.75	0.84	0.86
Stat Log Heart	0.77	0.88	0.89
Post Operative Patient	0.78	0.87	0.88

Таблиця 4. Порівняльні результати методів за характеристикою Rand Index

Вибірка	K-Means	K-Prototype	AFC_PCEP
Hepatitis	0.72	0.73	0.74
Cancer	0.53	0.56	0.62
Stat Log Heart	0.56	0.58	0.59
Post Operative Patient	0.41	0.45	0.48

Таблиця 5. Порівняльні результати методів за характеристикою Jaccard Index

Вибірка	K-Means	K-Prototype	AFC_PCEP
Hepatitis	0.62	0.63	0.65
Cancer	0.45	0.46	0.48
Stat Log Heart	0.54	0.56	0.71
Post Operative Patient	0.33	0.35	0.38

Таблиця 6. Порівняльні результати методів за ентропією

Вибірка	K-Means	K-Prototype	AFC_PCEP
Hepatitis	0.52	0.52	0.52
Cancer	0.45	0.43	0.45
Stat Log Heart	0.45	0.44	0.43
Post Operative Patient	0.42	0.41	0.40

Зменшення значень виміру ентропії свідчить про кращу продуктивність. Виходячи з цього, робота методу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур (AFC_PCEP) на основі ентропії значно вище, ніж K-Means і K-Prototypes для всього набору даних, що продемонстровано в таблиці 6.

Висновки

Розглянуто задачу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу - центроїда з використанням еволюційних процедур. Оскільки цільові функції нечіткої кластеризації в загальному випадку є багатоекстремальними, запропоновано уточнювати отримані центри кластерів за допомогою еволюційного методу котячих зграй. Запропоновано модифікацію, введену на основі процедури оптимізації котячих зграй з покращеними властивостями за рахунок використання стохастичної оцінки градієнта. Запропонований метод є простим у чисельній реалізації, працездатним у випадку, коли дані пошкоджені та надходять послідовно в online-режимі, що підтверджено експериментально.

References

1. Rutkowski, L. (2008) “Computational Intelligence Methods and Techniques”, Springer-Verlag, Berlin Heidelberg, 514 p.
2. Mumford, C., Jain, L. (2009) “Computational Intelligence. Collaboration, Fusion and Emergence”, Springer-Verlag, Berlin Heidelberg, 729 p.
3. Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., Held, P. (2013) “Computational Intelligence. A Methodological Introduction”, Springer, Berlin, 488 p.
4. Grosan, C., Abraham, A., Chis, M.: Swarm intelligence in Data Mining - Studies in Computational Intelligence (2006).
5. Chu, S.-C., Tsai, P.-W., Pan, J.S.: Cat swarm optimization. In: Lecture Notes in Artificial Intelligence. Berlin Heidelberg: Springer-Verlag (2006).
6. Chu, S.-C., Tsai, P.-W.: Computational Intelligence based on the behavior of cats.In: “Int. J. of Innovative Computing, Information, and Control”, №1, pp.163 – 173 (2007).

7. Kennedy, J., Eberhart, R. (1995) “Particle swarm optimization”, Proc. IEEE Int. Conf. on Neural Networks, Vol. 4, P. 1942 – 1948.
8. Chu, S.-C., Tsai, P.-W., Pan, J.S. (2006) “Cat swarm optimization”, Lecture Notes in Artifi-cial Intellgence,4099, Berlin Heidelberg, Springer-Verlag, P. 854-858.
9. Chu, S.-C., Tsai, P.-W. (2007) “Computational Intelligence based on the behavior of cats “, Int. J. of Innovative Computing, Information, and Control, 3, №1, pp. 163 – 173.
10. Shafronenko A., Bodyanskiy Ye., Rudenko (2019) “Online neuro fuzzy clustering of data with omissions and outliers based on comple-tion strategy“, Proceedings of The Second In-ternational Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), 2019. Zaporizhzhia, 2019. - P.18-27.
11. Shafronenko, A., Bodyanskiy, Ye. Pliss, I., Patlan, K.: Fuzzy Clusterization of Distorted by Missing Observations Data Sets Using Evo-lutionary Optimization. In: Proceedings “Advanced Computer Information Technologies (ACIT’2019)”, Česke Budejovice, Czech Republic, June 5-7, 2019, pp. 217-220 (2019). doi: 10.1109/ACITT.2019.8779888.
7. Shafronenko A., Bodyanskiy Ye., Klymova I., Holovin O. Online credibilistic fuzzy clustering of data using membership functions of special type. Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27-1 May 2020. - Zaporizhzhia, 2020.
8. Shafronenko A. Yu, Bodyanskiy Ye. V., Pliss I.P. The Fast Modification of Evolutionary Bioinspired Cat Swarm Optimization Method. Proc.2019 IEEE 8th International Conference on Advanced Optoe-lectronics and Lasers (CAOL), 2019. - Sozopol, Bul-garia, 2019. - P. 548-552. DOI: 10.1109 /CAOL46282. 2019.9019583.

Стаття надійшла до редакції 07.04.22
Після обробки 12.05.22